Learning for Visual Synthesis and Transformation

Xinyuan Chen

Faculty of Engineering and Information Technology

University of Technology Sydney

A thesis submitted for the degree of

Doctor of Philosophy

April 2020

I would like to dedicate this thesis to my loving parents
Lianfu Chen and *Meiyun Mao*

Certificate of Original Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for the collaborative degree and/or fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This research is supported by the Australian Government Research Training Program.

This thesis is the result of a Collaborative Doctoral Research Degreee program with Shanghai Jiao Tong University.

signature: Production Note: Signature removed prior to publication.

Acknowledgements

I would like to take this good opportunity to appreciate my advisors, several professors, my colleagues, my friends and my family for their significant help during my doctoral study in Shanghai Jiao Tong University.

My deepest gratitude goes first and foremost to my supervisor Prof. Dacheng Tao for his continuous support, unlimited patience and supportive guidance. He was always ready to help, and was very willing to teach everything he knows to me. His incredible enthusiasm and high demand motivate me to set up high academic research standards and submit papers to the leading journals or conferences in my research field.

I would also like to thank my supervisor Prof. Xiaokang Yang from Shanghai Jiao Tong University. He has given me not only plenty of freedom to explore in my research field, but also numerous constructive suggestions to help me out of various difficulties in study and life. I am particularly impressed by his encouraging attitude and expert knowledge for my research. Without his high academic standard, patient guidance, and unreserved support, I can never imagine I could have finished this thesis.

Besides my principal supervisors, I would like to give special thanks to my advisor: Dr. Chang Xu. I am grateful for his great patience in answering my simple and nearly endless questions from research motivation to implementation details, as well as improving my writing word-for-word. He taught me how to find interesting ideas, how to develop solid algorithms, and how to write technical papers all from scratch. I also thank Dr. Youming Qiao, Dr. Guoqiang Zhang

in UTS, Dr. Tongliang Liu, Dr. Shaoli Huang in USYD for their generous help during my study in Sydney.

I would also like to thank Prof. Li Song who led me into the field of computer vision. During my early days of academic research in Shanghai Jiao Tong University, he gave me valuable instructions and generous support. I also would like to thank Dr. Chao Ma. As a senior, he set up a great role for me. As an advisor, he provided me a lot of suggestions and help. Moreover, I would like to thank Prof. Bingbing Ni, Yi Xu and Guangtao Zhai for their advices and help.

The most enjoyable thing during my Ph.D. study is the opportunity to meet my dear colleagues and friends. I would like to thank Yilin Dong, Mengyue Shi, Yihang Huang, Han Zhang, Shaowei Xie, Zhe Ren, Genning Zhang for being close friends and offering me timely companionship; thank Muming Zhao, Dr. Jiangchao Yang, and Guo Lu for generous help and company in both Shanghai and Sydney; and thank Dr. Yichao Yan, Minsi Wang, Jingwei Xu, Wenbo Bao, Yunqian Wen, Jun Ling, Han Xue and many others.

Then comes my dear colleagues and friends in my doctoral study in Sydney: Dr. Erkun Yang, Dr. Chaoyue Wang, Yuxuan Du, Yali Du, Dr. Shan You, Dr. Jianfeng Dong, Yuangang Pan, Dr. Baosheng Yu, Dr. Liu Liu, Dr. Xiyu Yu, Dr. Huan Fu, Shanshan Zhao, Lianbo Zhang, Dalu Guo, Zeyu Feng, Zhe Chen, Rui Geng, Gengxing Wang and Xiaofei Liu. To my friends, thanks for all your support and company during my joyful and stressful time, and I cherish our friendship. To my colleague, it is fantastic to have the opportunity to work with you.

Finally, I want to dedicate this thesis to my parents who gave me unreserved love and support. Every Friday night, you always wait for my phone to hear my experience of study and life, share my joys and sorrows. Because of you, I never feel alone and frustrated through my difficult times. Thanks for always being there for me.

Abstract

Visual synthesis is one of the most fundamental problems in computer vision and artificial intelligence. Visual synthesis aims to create pixel-level data (e.g., images and videos) based on descriptions such as texts, noise, semantic annotations and images. Recently, deep generative learning has greatly promoted the development of visual synthesis. However, the existing generative methods still suffer from several issues, including model interpretation, controllability, stability, efficiency and performance. In this thesis, several generative models are proposed to address these challenges. This thesis makes the following contributions:

First, this thesis introduces an attention generative model for local image synthesis, so as to improve the controllability and interpretation of the generative model. How to precisely locate the foreground region in the image and generate the target object to the specified region is the key problem in the local image synthesis task. The object transfiguration task is an application of the local image synthesis, which aims to transform the object of images to another object. Existing generative methods often fail to decompose the foreground and background. In this thesis, the attention mechanism is incorporated into generative models, so as to transform the object of our interests without altering the background. The model is built by decomposing the generative network into two separate networks, each of which is dedicated to one sub-task: to detect the region of interests and to generate the object from one object to another. The attention network predicts spatial attention maps of images, and the transformation network focuses on translating objects. The attention network produces attention maps which are encouraged to be sparse so that the model

can only pay attention to the objects of interest. Also, a novel perceptual loss is introduced to improve the quality of transformed images in the high-level feature space. Experimental results demonstrate the necessity of investigating attention in image-to-image transformation, and that the improvement of the quality of generated images.

Second, this thesis proposes a multi-domain generative model that multiple styles of images can be generated in a single network. The major challenge is how to efficiently generate multiple styles in a single network. Our model is capable to extract the content and style feature of images, and apply multiple style features to the content image.This thesis proposes a gated generative model that consists of three modules: an encoder, a gated transformer, and a decoder. Different styles can be achieved through different branches of gated transformers while the encoder and decoder are used for capturing structure information sharing weights for all styles. A discriminative network is used to distinguish whether the input image is a stylized or genuine image. An auxiliary classifier is used to recognize the style categories of transferred images, thereby helping to generate images in multiple styles. In addition, to stabilize the adversarial training process, an auto-encoder reconstruction loss is introduced by combining the encoder and decoder module. Extensive experiments demonstrate the stability and effectiveness of the proposed model for multi-domain image synthesis.

Third, this thesis investigates the video synthesis problem on the long-term horizon. A temporal generative model is proposed for long-term video frame prediction. The existing generative model for video prediction usually cannot output high-quality predictions for a long-time horizon. The reason is that those methods recursively output subsequent frames by taking the newly generated frames as observations, consequently the prediction error accumulates dramatically. The introduced retrospection process is designed to look back on what has been learned from the past and rectify the prediction deficiencies.

To this end, a retrospection network is built to reconstruct the past frames given the currently predicted frames. On the other hand, an auxiliary route is built by reversing the flow of time and executing a similar retrospection. These two routes interact with each other to boost the performance of retrospection network and enhance the understanding of dynamics across frames, especially for the long-term horizon.

Overall, this thesis investigates the deep generative model and solves several practical issues for visual synthesis and transformation. For local image synthesis, we propose an attention generative model. We also propose a gated generative model for generating multi-domain of images in a single generative network. For video synthesis, a temporal generative model is proposed to output long-term video frames by incorporating the prediction and retrospection process in the model. Extensive experimental results on large-scale benchmark datasets demonstrate that the proposed methods in this thesis perform favorably against previous visual synthesis algorithms in terms of efficiency, controllability, and robustness.

Contents

List of Figures